BAYESIAN OUTPUT ANALYSIS PROGRAM (BOA) VERSION 1.0 USER'S MANUAL

Brian J. Smith

January 8, 2003

Contents

1 Getting Started **4**
- 1.1 Hardware/Software Requirements . . . 4
- 1.2 Obtaining BOA . . . 4
- 1.3 Installation . . . 4
- 1.4 BUGS Line Example . . . 5

2 Using the BOA Menu-Driven User Interface **6**

3 File Menu **8**
- 3.1 Import Data Menu . . . 8
  - 3.1.1 BUGS Output File . . . 9
  - 3.1.2 Flat ASCII File . . . 9
  - 3.1.3 Data Matrix Object . . . 10
  - 3.1.4 View Format Specifications . . . 10
  - 3.1.5 Data Options . . . 10
- 3.2 Load Session . . . 11
- 3.3 Save Session . . . 11
- 3.4 Exit BOA . . . 11

4 Data Management Menu **12**
- 4.1 Chains Menu . . . 12
  - 4.1.1 Combine All Chains . . . 13
  - 4.1.2 Delete Chain . . . 13
  - 4.1.3 Subset Chains . . . 14
- 4.2 Parameters Menu . . . 15
  - 4.2.1 Set Parameter Bounds . . . 15
  - 4.2.2 Delete Parameters . . . 16
  - 4.2.3 Create New Parameters . . . 17

4.3 Display Working Dataset . . . 17
4.4 Display Master Dataset . . . 18
4.5 Reset . . . 19

5 Analysis Menu **20**
5.1 Descriptive Statistics Menu . . . 20
5.1.1 Autocorrelations . . . 21
5.1.2 Correlation Matrix . . . 21
5.1.3 Highest Probability Density Intervals . . . 22
5.1.4 Summary Statistic . . . 22
5.2 Convergence Diagnostics Menu . . . 23
5.2.1 Brooks, Gelman & Rubin Convergence Diagnostic . . . 24
5.2.2 Geweke Convergence Diagnostic . . . 25
5.2.3 Heidelberger and Welch Convergence Diagnostic . . . 26
5.2.4 Raftery and Lewis Convergence Diagnostic . . . 27
5.3 Analysis Options . . . 28

6 Plot Menu **30**
6.1 Descriptive Plot Menu . . . 30
6.1.1 Autocorrelations Plot . . . 30
6.1.2 Density Plot . . . 31
6.1.3 Running Mean Plot . . . 31
6.1.4 Trace Plot . . . 31
6.2 Convergence Diagnostics Plot Menu . . . 32
6.2.1 Brooks and Gelman Plot . . . 32
6.2.2 Gelman and Rubin Plot . . . 33
6.2.3 Geweke Plot . . . 34
6.3 Plot Options . . . 34

7 Options Menu **38**

8 Window Menu **39**
8.1 Previous Graphics Window . . . 39
8.2 Next Graphics Window . . . 39
8.3 Save to Postscript File . . . 40
8.4 Close Graphics Window . . . 40
8.5 Close All Graphics Window . . . 40

9 S-PLUS and R Basics **41**
9.1 Output Display Options 41
9.2 Vectors in S 41
9.3 Memory Allocation in R 42

Chapter 1

Getting Started

1.1 Hardware/Software Requirements

BOA has been successfully tested on S-PLUS versions 3.4-6.0 for UNIX, S-PLUS 2000 for Microsoft Windows, and R 1.2.0 for UNIX and Microsoft Windows.

1.2 Obtaining BOA

The program files and help documentation are freely available from http://www.public-health.uiowa.edu

1.3 Installation

The program source is located in the ASCII text file

boa.s (S-PLUS version) OR boa.r (R version)

UNIX S-PLUS and R users should source the appropriate file via

source("<program directory>/boa.s") OR source("<program directory>")

where <program directory> is the directory in which the program files are located. S-PLUS Windows users may open this as an S-PLUS script file and run that file from the script window. The BOA program need not be installed to the directory in which the data to be analyzed are located. The location of the data can be specified from within the program.

1.4 BUGS Line Example

Output from the BUGS Line example is used to illustrate the capabilities of the BOA program. The Line example involves a liner regression analysis of the data points (1, 1), (2, 3), (3, 3), (4, 3), and (5, 5). The proposed Bayesian model is

$$\begin{aligned} y[i] &\sim N(mu[i], tau) \\ mu[i] &= alpha + beta * (x[i] - mean(x[])) \end{aligned}$$

with the following priors:

$$\begin{aligned} alpha &\sim N(0, 0.0001) \\ beta &\sim N(0, 0.0001) \\ tau &\sim Gamma(0.001, 0.001) \end{aligned}$$

Interest lies in estimating the posterior distribution of *alpha*, *beta*, and $sigma = 1/sqrt(tau)$. The starting values for the parameters were varied to generate two parallel chains from the Markov chain Monte Carlo (MCMC) sampler. The first chain, line1, was generated with the initial values of

$$\text{list}(tau = 5, alpha = \text{-}5, beta = 5, \text{seed} = 987654321)$$

whereas, the second chain, line2, was generated with

$$\text{list}(tau = 0.01, alpha = 0.01, beta = 0.01, \text{seed} = 1234567890)$$

Each chain contains 200 iterations. The resulting output is available from the BOA website.

Chapter 2

Using the BOA Menu-Driven User Interface

A menu-driven interface is supplied with the BOA. It provides easy access to all of the command line function. To start the menu system, type

```
Bayesian Output Analysis Program (BOA)
Version 1.0.0 for UNIX R
Copyright (c) 2001 Brian J. Smith <brian-j-smith@uiowa.edu>

This program is free software; you can redistribute it and/or
modify it under the terms of the GNU General Public License
as published by the Free Software Foundation; either version 2
of the License or any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

For a copy of the GNU General Public License write to the Free
Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
MA  02111-1307, USA, or visit their web site at
http://www.gnu.org/copyleft/gpl.html

NOTE: if the menu unexpectedly terminates, type "boa.menu(recover=TRUE)"
to restart and recover your work
```

```
BOA MAIN MENU
*************
1:File     >>
2:Data     >>
3:Analysis >>
4:Plot     >>
5:Options  >>
6:Window   >>
Selection:
```

Note the message given at startup: if the menu unexpectedly terminates, type "boa.menu(recover = TRUE)" to restart and recover your work. There are a few instances where supplying the wrong type of data will crash the menu system. Immediately doing a recover will ensure that no data is lost.

Chapter 3

File Menu

Selecting menu item 1 from the BOA Main Menu brings up the File Menu. Options to import data, load previously saved session data, save the current session, and exit the program are available from the File Menu:

```
FILE MENU
=========
1:Import Data         >>
2:Load Session
3:Save Session
4:Return to Main Menu
5:Exit BOA
Selection:
```

3.1 Import Data Menu

BOA can import MCMC output from a variety of sources. Data may be imported at any point in the analysis from three different sources.

```
IMPORT DATA MENU
----------------
1:BUGS Output File
2:Flat ASCII File
3:Data Matrix Object
4:View Format Specifications
5:Options...
6:Back
```

```
7:Return to Main Menu
Selection:
```

3.1.1 BUGS Output File

The two CODA output files generated by the Bayesian inference Using Gibbs Sampling (BUGS or WinBUGS) program can be imported into BOA. The output file containing the parameter definitions should be saved as a .ind file; whereas, the file containing the sampler output should be saved as a .out file. BOA will expect these files to be located in the Working Directory. See Section 3.1.5 for instructions on specifying the working directory. Upon selecting to import BUGS output the user will be prompted to

```
Enter filename prefix without the .ind/.out extension [Working
Directory: "d:/bjsmith/boa"]
1: line1
```

Only the filename prefix should be specified. BOA will automatically add the appropriate extensions and load the data from the line1.ind and line1.out files.

3.1.2 Flat ASCII File

BOA includes an import filter for general ASCII files. This is particularly useful for output generated by custom MCMC programs. The ASCII file should contain one run of the sampler with the monitored parameters stored in space, comma, or tab delimited columns and the parameter names in the first row. Iteration numbers may be specified in a column labeled "iter". The ASCII file should be located in the Working Directory. Upon selected to import an ASCII file the program will prompt the user to

```
Enter filename prefix without the .txt extension [Working Directory:
"d:/bjsmith/boa/"]
1: line1
```

Specify only the filename prefix. The import filter will automatically add the extension and load the data from the line1.txt file. See Section 3.1.5 for instructions on specifying the Working Directory and the default ASCII file extension.

3.1.3 Data Matrix Object

MCMC output stored as an S object may be imported into BOA. The object must be a numeric matrix whose columns contain the monitored parameters from one run of the sampler. The iteration numbers and parameter names may be specified in the dimnames. Upon selecting to import a matrix object the user will be asked to

```
Enter object name [none]
1: line1
```

BOA will import the data from the line1 object in the current S-PLUS or R session.

3.1.4 View Format Specifications

Selecting this menu item will display the format specifications for the three types of data that BOA can import.

```
BUGS
- Bayesian inference Using Gibbs Sampling output files (*.ind and *.out)
- files must be located in the Working Directory (see Options)

ASCII
- ASCII file (*.txt) containing the monitored parameters from one run of
the sampler
- file must be located in the Working Directory (see Options)
- parameters are stored in space, comma, or tab delimited columns
- parameter names must appear in the first row
- iteration numbers may be specified in a column labeled 'iter'

Matrix Object
- S or R numeric matrix whose columns contain the monitored parameters
from one run of the sampler
- iteration numbers and parameter names may be specified in the dimnames
```

3.1.5 Data Options

The Options... menu item lists the values for the user settings used to import data.

```
Data Parameters
===============
```

```
Files
-----
1) Working Directory: "/home/bjsmith/";
2) ASCII File Ext:    ".txt"

Select parameter to change or press <ENTER> to continue
1:
```

Most users will want to specify the Working Directory at the start of their BOA session. This directory should be set to the path in which the MCMC output files are stored. Forward slashes ("/") should be used when entering the directory path.

3.2 Load Session

The Load Session menu item allows users to load previously saved work.

```
Enter name of object to load [none]
1: line
```

3.3 Save Session

All imported data and user settings may be saved at any point in the analysis. Users will be prompted to

```
Enter name of object to which to save the session data [none]
1: line
```

The session data will be saved to the specified S object.

3.4 Exit BOA

Select this item to exit from the BOA program. Users will be prompted to verify their intention to exit in order to avoid an unintended termination of the program.

```
Do you really want to EXIT (y/n) [n]?
1:
```

Users wishing to save their work should go back and do so before exiting. BOA will not automatically save the user's work.

Chapter 4

Data Management Menu

BOA offers a wide range of options for managing the imported data. Two copies of the data are maintained by the program - the Master dataset and the Working dataset. The Master dataset is a static copy of the data as it was first imported. This copy remains essentially unchanged throughout the BOA session. The Working dataset is a dynamic copy that can be modified by the user. All analyses are performed on the Working dataset. The Data Management menu offers the following options:

```
DATA MANAGEMENT MENU
====================
1:Chains                  >>
2:Parameters              >>
3:Display Working Dataset
4:Display Master Dataset
5:Reset
6:Return to Main Menu
Selection:
```

4.1 Chains Menu

```
CHAINS MENU
-----------
1:Combine All
2:Delete
3:Subset
4:Back
```

```
5:Return to Main Menu
Selection:
```

4.1.1 Combine All Chains

Selecting this options will combine together all of the chains in the Working dataset. Sequencing is preserved by concatenating together the different chains and then ordering the result by the iteration numbers in the original chains. Note that this may result in a chain with multiple samples at a given iteration. The resulting chain contains only those parameters common to all chains.

CAUTION: Although possible to do so, convergence diagnostics and autocorrelations should not be computed for combined chains. A combined chain is essentially a single chain with potentially multiple samples per iteration. These analyses expect that a single chain has no more than one sample per iteration.

4.1.2 Delete Chain

Chains may be discarded when they are no longer needed. Discarding chains may free up a substantial amount of computer memory. The program prompts the user to select the chain(s) to discard.

```
DISCARD CHAINS
==============

Chains:
-------

      1       2
"line1" "line2"

Specify chain index or vector of indices [none]
1:
```

The specified chain(s) will be immediately deleted from the Master dataset. If the Working dataset has not been modified, the chain(s) will be deleted from there as well. If modifications were made to the Working dataset, the user can copy the new Master dataset to the Working dataset via the Reset option. If no chain is entered at the prompt, no action is taken.

4.1.3 Subset Chains

Subsets of the MCMC sequences can be selected for analysis via the Subset option.

```
SUBSET WORKING DATASET
======================
Specify the indices of the items to be included in the subset.
Alternatively, items may be excluded by supplying negative indices.
Selections should be in the form of a number or numeric vector.

Chains:
-------

      1       2
"line1" "line2"

Specify chain indices [all]
1: c(1,2)

Parameters:
-----------

      1      2       3
"alpha" "beta" "sigma"

Specify parameter indices [all]
1: -2

Iterations:
++++++++++

      Min Max Sample
line1   1 200    200
line2   1 200    200

Specify iterations [all]
1: 50:200
```

In this example, both chains were first included in the subset. Since the default is to include all chains, this line could have been left blank. Next, the *beta* parameter is excluded by supplying a negative sign in front of the selection. Finally, the

subset is limited to iteration 50-200. Users can verify that the subset was successfully constructed by selecting the option to display the Working dataset (output not shown).

Thinning: Thinning refers to the practice of including every k^{th} iteration from a chain. Users can thin a chain by using the *seq* function when prompted to specify the iterations. For example, the following input will included every other iteration from the chain:

```
seq(1, 200, length = 100)
```

A description of the *seq* function can be found at the end of the Appendix.

4.2 Parameters Menu

```
PARAMETERS MENU
---------------
1:Set Bounds
2:Delete
3:New
4:Back
5:Return to Main Menu
Selection:
```

4.2.1 Set Parameter Bounds

This option allows the user to specify the lower and upper bounds (support) of selected MCMC parameters. The parameter support factors into the computation of the Brooks, Gelman & Rubin convergence diagnostics.

```
SET PARAMETER BOUNDS
====================

Chains:
-------

      1       2
"line1" "line2"
```

```
Specify chain index or vector of indices [all]
1:

Parameters:
-----------

      1       2       3
"alpha"  "beta" "sigma"

Specify parameter index or vector of indices [all]
1: 3

Specify lower and upper bounds as a vector [-Inf, Inf]
1: c(0, Inf)
```

In this example, the variance parameter *sigma* has been restricted to only non-negative values. When no chain(s) is specified, the default is to apply the change to all of the chains. Likewise, the default is to select all parameters and to set the bounds to (-Infinity, Infinity).

4.2.2 Delete Parameters

Often times it may be desired to delete parameters that are not of interest in the analysis. This may arise in cases where data other than model parameters were saved to the output file imported into BOA. Alternatively, the user may only be interested in functions of the original parameters. Once the new parameter is created using the methods described in the following section, the unnecessary parameter upon which it is based may be deleted. Deleted parameters will speed up the manipulation of data in BOA.

```
DELETE PARAMETERS
=================

Parameters:
-----------

      1       2       3
"alpha"  "beta" "sigma"

Specify parameter index or vector of indices [none]
```

4.2.3 Create New Parameters

BOA includes an option to create new parameters. Most S functions can be used to create the new parameter. Typically, a new parameter is defined as a function of the existing parameters. For instance, suppose the user was interested in analyzing the precision parameter $tau = 1/sigma^2$. The following menu commands demonstrate how to create this precision parameter:

```
NEW PARAMETER
=============

Common Parameters:
------------------

[1] "alpha" "beta"  "sigma"

New parameter name [none]
1: tau
Read 1 items

Define the new parameter as a function of the parameters listed above
1: 1 / sigma^2
Read 1 items
```

tau has now been added to the two datasets in BOA and will be available to all subsequent analyses.

4.3 Display Working Dataset

Selecting this option will display summary information for the Working dataset.

```
CHAIN SUMMARY INFORMATION:
==========================

Iterations:
++++++++++++

      Min Max Sample
line1   1 200    200
line2   1 200    200
```

```
Support: line1
--------------

    alpha beta sigma tau
Min  -Inf -Inf     0   0
Max   Inf  Inf   Inf Inf

Support: line2
--------------

    alpha beta sigma tau
Min  -Inf -Inf     0   0
Max   Inf  Inf   Inf Inf
```

In this example, the two chains have been combined. Hence, the Working dataset is a modified version of the Master dataset.

4.4 Display Master Dataset

Selecting this option will display summary information for the Master dataset.

```
CHAIN SUMMARY INFORMATION:
==========================

Iterations:
++++++++++++

      Min Max Sample
line1   1 200    200
line2   1 200    200

Support: line1
--------------

    alpha beta sigma tau
Min  -Inf -Inf     0   0
Max   Inf  Inf   Inf Inf
```

```
Support: line2
--------------

    alpha beta sigma tau
Min  -Inf -Inf     0   0
Max   Inf  Inf   Inf Inf
```

4.5 Reset

The Reset option copies the Master dataset to the Working dataset. This undoes any modifications that were made to the Working dataset.

Chapter 5

Analysis Menu

The statistical analysis procedures are accessible through the Analysis Menu. Analyses are categorized into two groups – Descriptive Statistics and Convergence Diagnostics.

```
ANALYSIS MENU
=============
1: Descriptive Statistics  >>
2: Convergence Diagnostics >>
3: Options...
4: Return to Main Menu
Selection:
```

5.1 Descriptive Statistics Menu

Options to compute autocorrelations, cross-correlations, and summary statistics are available from the Descriptive Statistics Menu.

```
DESCRIPTIVE STATISTICS MENU
---------------------------
1:Autocorrelations
2:Correlation Matrix
3:Highest Probability Density Intervals
4:Summary Statistics
5:Back
6:Return to Main Menu
Selection:
```

5.1.1 Autocorrelations

This option produces lag-autocorrelations for the monitored parameters within each chain. High autocorrelations indicate slow mixing within a chain and, usually, slow convergence to the posterior distribution.

```
LAGS AND AUTOCORRELATIONS:
==========================

Chain: line1
------------

          Lag 1      Lag 5     Lag 10     Lag 50
alpha -0.141244 -0.0935154 -0.1139290  0.1150173
beta   0.240793  0.2157423  0.1174318 -0.0240399
sigma  0.328217  0.1068319 -0.0015778 -0.0292616
tau    0.296238  0.1474137  0.0114918 -0.0073957
```

Option 11 in Section 5.3 allows the user to set the lags at which autocorrelations are computed.

5.1.2 Correlation Matrix

This option returns the correlation matrix for the parameters in each chain. High correlation among parameters may lead to slow convergence to the posterior. Associated models may need to be reparameterized in order to reduce the amount of cross-correlation.

```
CROSS-CORRELATION MATRIX:
=========================

Chain: line1
------------

          alpha      beta     sigma  tau
alpha         1
beta  -0.205461         1
sigma  0.176248 -0.312405         1
tau   -0.034745  0.052883 -0.468271    1
```

5.1.3 Highest Probability Density Intervals

Highest probability density (HPD) interval estimation is one means of generating Bayesian posterior intervals. HPD intervals span a region of values containing $(1 - \alpha) \times 100\%$ of the posterior density such that the posterior density within the interval is always greater than that outside. Consequently, HPD intervals are of the shortest length of any of the Bayesian intervals. The algorithm described by Chen and Shao (1999) is used to compute the HPD intervals in BOA under the assumption of unimodal marginal posterior distributions. The alpha level for the intervals can be modified through Option 12 in Section 5.3.

```
HIGHEST PROBABILITY DENSITY INTERVALS:
======================================

Alpha level = 0.05

Chain: line1
------------

      Lower Bound Upper Bound
alpha  1.64596000     3.95041
beta  -0.17817400     1.52666
sigma  0.37893400     2.46407
tau    0.00575839     4.45471
```

5.1.4 Summary Statistic

This option prints summary statistics for the parameters in each chain. The sample mean and standard deviation are given in the first two columns. These are followed by three separate estimates of the standard error: 1) a naive estimate (the sample standard deviation divided by the square root of the sample size) which assumes the sampled values are independent, 2) a time–series estimate (the square root of the spectral density variance estimate divided by the sample size) which gives the asymptotic standard error (Geweke, 1992), and 3) a batch estimate calculated as the sample standard deviation of the means from consecutive batches of size 50 divided by the square root of the number of batches. The autocorrelation between batch means follows and should be close to zero. If not, the batch size should be increased. Quantiles are given after the batch autocorrelation. Finally, the minimum and maximum iteration numbers and the total sample size round out the table.

```
SUMMARY STATISTICS:
===================

Batch size for calculating Batch SE and (Lag 1) ACF = 50

Chain: line1
------------

          Mean       SD  Naive SE  MC Error  Batch SE  Batch ACF
alpha 3.012561 0.597966 0.0422826 0.0308999 0.0242799  0.1413237
beta  0.790924 0.444957 0.0314632 0.0280210 0.0693133 -0.0971101
sigma 1.182538 1.212897 0.0857648 0.0567591 0.1184687 -0.3510225
tau   1.726886 1.507967 0.1066293 0.1274516 0.1306512 -0.4207326

          0.025     0.5   0.975 MinIter MaxIter Sample
alpha  1.748050 2.98737 4.07063       1     200    200
beta  -0.107853 0.81173 1.63205       1     200    200
sigma  0.458036 0.86806 3.28955       1     200    200
tau    0.092636 1.32741 4.76732       1     200    200
```

Options 13 and 14 in Section 5.3 allow the user to change the batch size and the quantiles, respectively. See the Appendix for instructions on setting the number of significant digits and display width.

5.2 Convergence Diagnostics Menu

The Convergence Diagnostics Menu offers the user the following diagnostic methods:

```
CONVERGENCE DIAGNOSTICS MENU
----------------------------
1: Brooks, Gelman & Rubin
2: Geweke
3: Heidelberger & Welch
4: Raftery & Lewis
5: Back
6: Return to Main Menu
Selection:
```

These are the most commonly used methods used to asses the convergence of MCMC output. A brief explanation of each approach is given in the following sections. Users

are referred to the work of Brooks and Roberts (1998) and Cowles and Carlin (1996) for a more in-depth review and comparison of these methods.

5.2.1 Brooks, Gelman & Rubin Convergence Diagnostic

The code for implementing the Gelman and Rubin (1992) convergence diagnostic in BOA is based on the *itsim* function contributed to the Statlib archive by Andrew Gelman (http://lib.stat.cmu.edu).

The Brooks, Gelman and Rubin convergence diagnostic is appropriate for the analysis of two or more parallel chains, each with different starting values which are overdispersed with respect to the target distribution. Several methods for generating starting values for the MCMC samplers have been proposed (Gelman and Rubin, 1992; Applegate et al., 1990; Jennison, 1993). The following diagnostic information was obtained for the line example:

```
BROOKS, GELMAN AND RUBIN CONVERGENCE DIAGNOSTICS:
================================================

Iterations used = 101:200

Potential Scale Reduction Factors
---------------------------------

   alpha    beta   sigma
 1.00316 1.00171 1.00841

Multivariate Potential Scale Reduction Factor = 1.026086

Corrected Scale Reduction Factors
---------------------------------

      Estimate   0.975
alpha  1.00596 1.03859
beta   1.00940 1.03702
sigma  1.01147 1.06440
```

The diagnostic originally proposed by Gelman and Rubin (1992) is based on a comparison of the within and between chain variance for each variable. This comparison is used to estimate the *potential scale reduction factor* (PSRF) – the multiplicative

factor by which the sampling-based estimate of the scale parameter of the marginal posterior distribution might be reduced if the chains were run to infinity. To adjust for the sampling variability in the variance estimates, the correction proposed by Brooks and Gelman (1998) is applied to the PSRF to produce the *corrected scale reduction factor* (CSRF). BOA also displays an upper quantile of the sampling distribution for the CSRF. Users can control which quantile is computed via Option 1 in Section 5.3. Brooks and Gelman (1998) developed a multivariate extension to the PSRF known as the *multivariate potential scale reduction factor* (MPSRF). The MPSRF does not include a correction for sampling variability. This statistic is relevant when interest lies in general multivariate functionals of the chain. The MPSRF and the PSRF satisfy the following relationship:

$$max(\text{PSRF}) \leq \text{MPSRF}$$

Computation of the reduction factors is based on analysis of variance and sampling from the normal distribution. To avoid violations of the latter assumption, BOA transforms any parameters specified to be restricted to the range (a, b) to the logarithmic or logit scale before calculating this diagnostic. By default only the second half of the chains (iterations 101-200) is used to compute the reduction factors. Option 2 in 5.3 can be used to vary the proportion of samples from the end of the chains to be included in the analysis. If the estimates are approximately equal to one (or, as a rule of thumb, the 0.975 quantile is ≤ 1.2), the samples may be considered to have arisen from the stationary distribution. In this case, descriptive statistics may be calculated for the combined latter 50% of iterations from all of the chains.

5.2.2 Geweke Convergence Diagnostic

The Geweke convergence diagnostic is appropriate for the analysis of individual chains when convergence of the *mean* of some function of the sampled parameters is of interest. The following diagnostic information was obtained for the line example:

```
GEWEKE CONVERGENCE DIAGNOSTIC:
==============================

Fraction in first window = 0.1
Fraction in last window = 0.5

Chain: line1
------------
```

```
           alpha         beta      sigma
Z-Score 0.251456 -3.338157835 3.07513466
p-value 0.801462  0.000843358 0.00210408
```

The chain is divided into two "windows" containing a set fraction of the first and the last iterations. Options 3 and 4 in Section 5.3 allow the user to set the fraction of iterations included in the first and the last window, respectively. Geweke (1992) proposed a method to compare the mean of the sampled values in the first window to the mean of the sampled values in the last window. There should be a sufficient number of iterations between the two windows to reasonably assume that the two means are approximately independent. His method produces a Z statistic calculated as the difference between the two means divided by the asymptotic standard error of their difference, where the variance is determined by spectral density estimation. As the number of iterations approaches infinity, the Z statistic approaches the $N(0,1)$ if the chain has converged. Z values which fall in the extreme tails of the $N(0,1)$ suggest that the chain in the first window had not fully converged. The two-sided p-value outputted by BOA gives the tail probability associated with the observed Z statistic. It is common practice to conclude that there is evidence against convergence when the p-value is less than 0.05. Otherwise, it can be said that the results of this test do not provide any evidence against convergence. This does not, however, prove that the chain has converged.

5.2.3 Heidelberger and Welch Convergence Diagnostic

The Heidelberger and Welch convergence diagnostic is appropriate for the analysis of individual chains. The following diagnostic information was obtained for the line example:

```
HEIDLEBERGER AND WELCH STATIONARITY AND INTERVAL HALFWIDTH TESTS:
================================================================

Halfwidth test accuracy = 0.1

Chain: line1
------------

      Stationarity Test Keep Discard  C-von-M Halfwidth Test    Mean
alpha            passed  200       0 0.208584         passed 3.01256
```

```
beta             passed  180      20 0.267382          passed 0.83503
sigma            passed  180      20 0.391229          passed 1.05928

      Halfwidth
alpha 0.0605627
beta  0.0561035
sigma 0.1026402
```

Heidelberger and Welch's (1983) stationarity test is based on Brownian bridge theory and uses the Cramer-von-Mises statistic. If there is evidence of non-stationarity, the test is repeated after discarding the first 10% of the iterations. This process continues until the resulting chain passes the test or more than 50% of the iterations have been discarded. BOA reports the number of iterations that were kept, the number of iterations that were discarded, and the Cramer-von-Mises statistic. Failure of the chain to pass this test indicates that a longer run of the MCMC sampler is needed in order to achieve convergence.

A halfwidth test is performed on the portion of the chain that passes the stationarity test for each variable. Spectral density estimation is used to compute the asymptotic standard error of the mean. If the halfwidth of the confidence interval for the mean is less than a specified fraction (accuracy) of this mean, the halfwidth test indicates that the mean is estimated with acceptable accuracy. The confidence level and accuracy can be modified through Options 5 and 6, respectively, in Section 5.3. Failure of the halfwidth test implies that a longer run of the MCMC sampler is needed to increase the accuracy of the estimated posterior mean.

5.2.4 Raftery and Lewis Convergence Diagnostic

The Raftery and Lewis convergence diagnostic is appropriate for the analysis of individual chains. The following diagnostic information was obtained for the line example:

```
RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:
=========================================

Quantile = 0.025
Accuracy = +/- 0.02
Probability = 0.9

Chain: line1
------------
```

```
       Thin Burn-in Total Lower Bound Dependence Factor
alpha     1       2   160         165           0.969697
beta      1       5   188         165           1.139394
sigma     1       2   160         165           0.969697
```

The diagnostic proposed by Raftery and Lewis (1992b) tests for convergence to the stationary distribution and estimates the run-lengths needed to accurately estimate quantiles of functions of the parameters. The user may specify the quantile of interest, the desired degree of accuracy in estimating this quantile, and the probability of attaining the indicated degree of accuracy. Options 7, 9, and 10 in Section 5.3 allow the user to modify these quantities. BOA computes the "lower bound" – the number of iterations needed to estimate the specified quantile to the desired accuracy using independent samples. If fewer iterations than this bound have been loaded into BOA, the following warning is displayed:

```
******* Warning *******
Available chain length is 200.
Re-run simulation for at least 3746 iterations
OR reduce the quantile, accuracy, or probability to be estimated.
```

If sufficient MCMC iterations are available, BOA lists the lower bound, the total number of iterations needed for each parameter, the number of initial iterations to discard as the burn-in set, and the thinning interval to be used. The dependence factor measures the multiplicative increase in the number of iterations needed to reach convergence due to within-chain correlation. Dependence factors greater than 5.0 often indicate convergence failure and a need to reparameterize the model (Raftery and Lewis, 1992a).

5.3 Analysis Options

```
Analysis Parameters
===================

Brooks, Gelman & Rubin
----------------------
1)  Alpha Level:       0.05
2)  Window Fraction:   0.5
```

```
Geweke
------
3)  Window 1 Fraction: 0.1
4)  Window 2 Fraction: 0.5

Heidelberger & Welch
--------------------
5)  Accuracy:          0.1
6)  Alpha Level:       0.05

Raftery & Lewis
---------------
7)  Accuracy:          0.005
8)  Alpha Level:       0.05
9)  Delta:             0.001
10) Quantile:          0.025

Statistics
----------
11) ACF Lags:          c(1, 5, 10, 50)
12) Alpha Level:       0.05
13) Batch Size:        50
14) Quantiles:         c(0.025, 0.5, 0.975)
```

Chapter 6

Plot Menu

Like the Analysis Menu, the Plot Menu categorizes the available plots into a Descriptive and Convergence Diagnostic group. Most of the options found under the Analysis Menu have a counterpart within the Plot Menu.

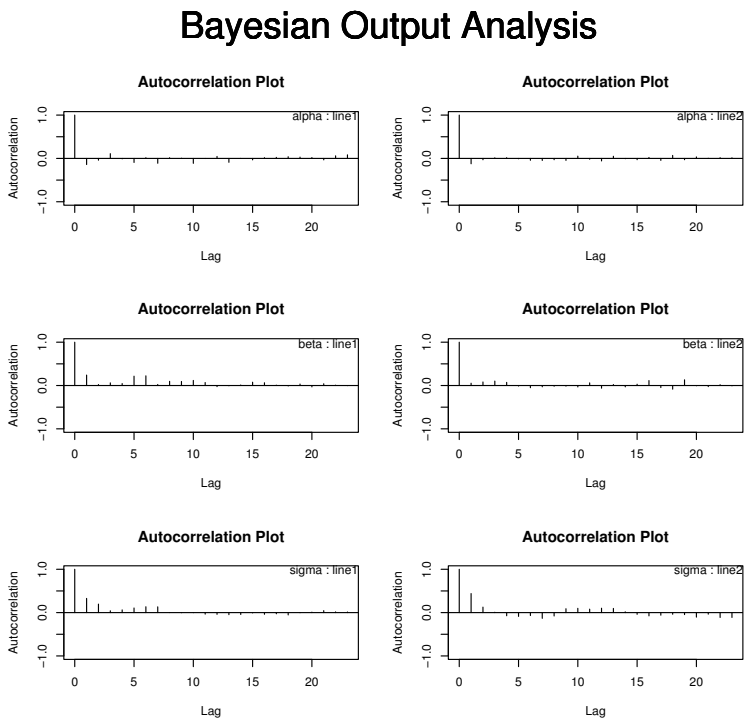
```
PLOT MENU
=========
1: Descriptive             >>
2: Convergence Diagnostics >>
3: Options...
4: Return to Main Menu
```

6.1 Descriptive Plot Menu

```
DESCRIPTIVE PLOT MENU
---------------------
1: Autocorrelations
2: Density
3: Running Mean
4: Trace
5: Back
6: Return to Main Menu
Selection:
```

6.1.1 Autocorrelations Plot

Plot the first several lag-autocorrelations for each parameter in each chain.



6.1.2 Density Plot

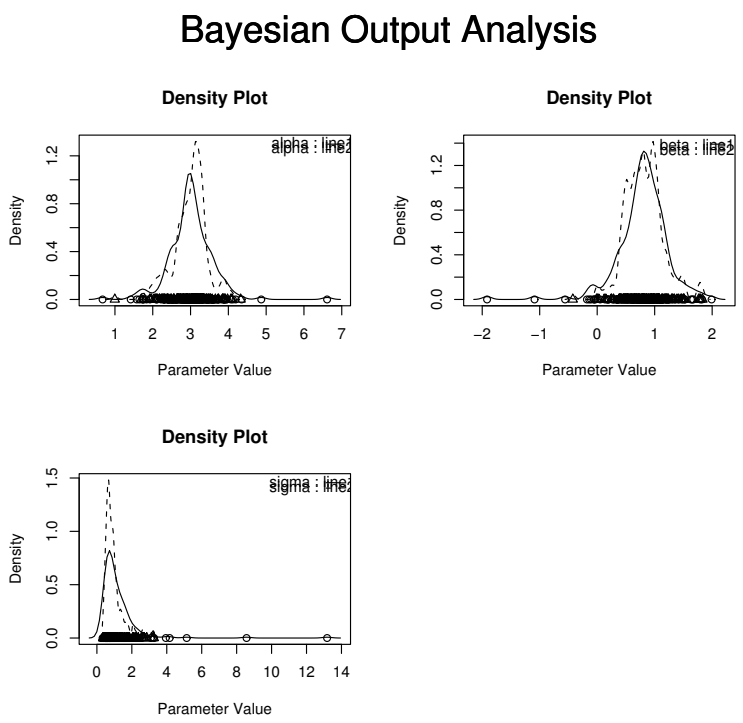
Plot the kernel density estimate for each parameters in each chain. Options 3 and 4 in Section 6.3 control the width and type of window used in the computations, respectively.

6.1.3 Running Mean Plot

Generate a time series plot of the running mean for each parameter in each chain. The running mean is computed as the mean of all sampled values up to and including that at a given iteration.

6.1.4 Trace Plot

Generate a time series plot of the sampled points for each parameter in each chain.

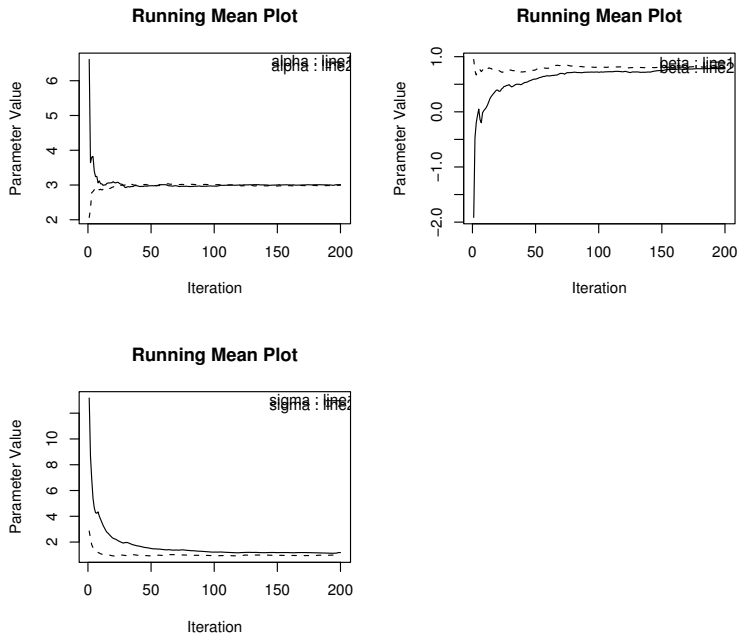


6.2 Convergence Diagnostics Plot Menu

```
CONVERGENCE DIAGNOSTICS PLOT MENU
---------------------------------
1: Brooks & Gelman
2: Gelman & Rubin
3: Geweke
4: Back
5: Return to Main Menu
Selection:
```

6.2.1 Brooks and Gelman Plot

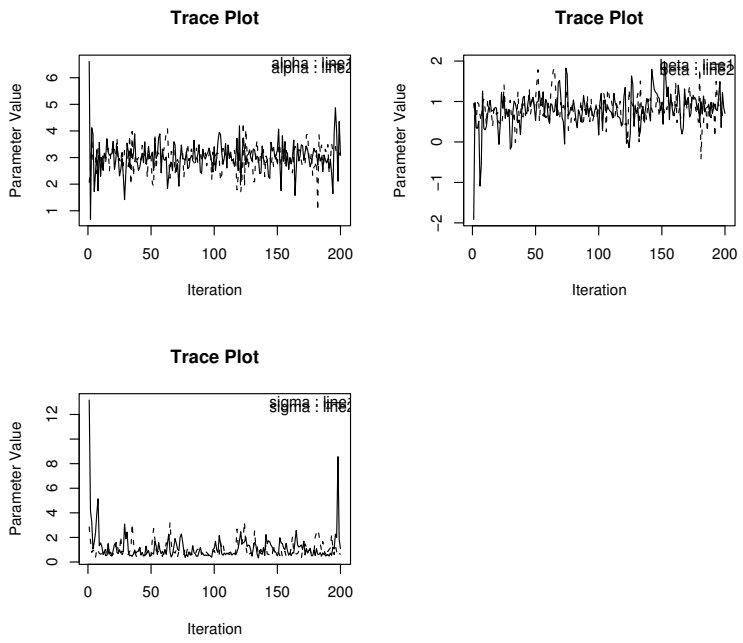
Plots the Brooks and Gelman *multivariate potential scale reduction factor* and the maximum of the *potential scale reduction factors* (see Section 5.2.1) for successively larger segments of the chains. The first segment contains the first 50 iterations in



the chains. The remaining iterations are then partitioned into equal bins and added incrementally to construct the remaining segments. Option 1 in Section 6.3 governs the number of bins used for the plot. Scale factors are plotted against the maximum iteration number in the segments. Cubic splines are used to interpolate through the point estimates from the segments.

6.2.2 Gelman and Rubin Plot

Plots the Gelman and Rubin *corrected potential scale reduction factors* (see Section 5.2.1) for each parameter in successively larger segments of the chain. The first segment contains the first 50 iterations in the chain. The remaining iterations are then partitioned into equal bins and added incrementally to construct the remaining segments. Options 5 and 6 in Section 6.3 control the error rate for the upper quantile and the number of bins, respectively. Option 7 determines the proportion of samples from the end of the chains to be included in the analysis. The scale factor is plotted against the maximum iteration number for the segment. Cubic splines are used to

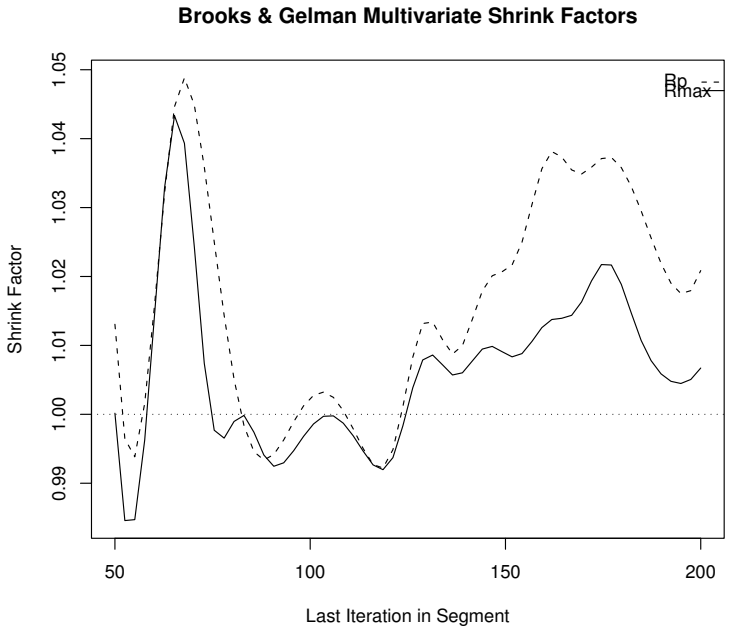


interpolate through the point estimates from the segments.

6.2.3 Geweke Plot

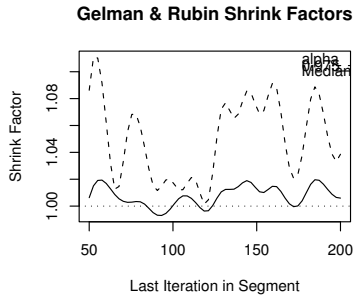
Plots the Geweke Z statistic (see Section 6.3) for each parameter in successively smaller segments of the chain. The k^{th} segment contains the last ((number of bins - k + 1) / number of bins)*100% of the iterations in the chain. Options 8 and 9 in Section 6.3 set the error rate for the confidence bounds and the number of bins included in the plot, respectively. Options 10 and 11 control the fraction of iterations covered by the windows used in computing the Geweke diagnostic. It may be possible that some of the subsets contain too few iterations to compute the test statistic. Such segments, if they exist, are automatically omitted from the plot. The test statistic is plotted against the minimum iteration number for the segment.

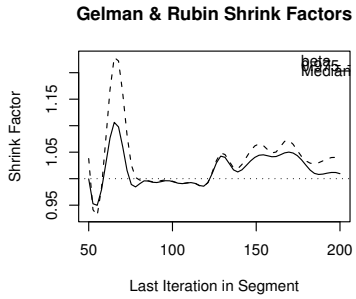
6.3 Plot Options

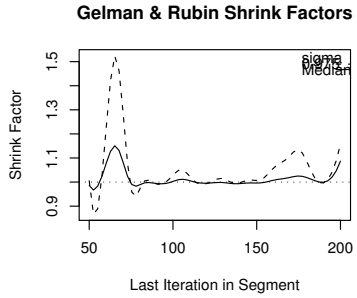



```
Plot Parameters
===============

Brooks & Gelman
---------------
1)  Number of Bins:      20
2)  Window Fraction:     0.5

Density
-------
3)  Bandwidth:           function (x)
                        0.5 * diff(range(x))/(log(length(x)) + 1)
4)  Kernel:              "gaussian"

Gelman & Rubin
--------------
5)  Alpha Level:         0.05
```

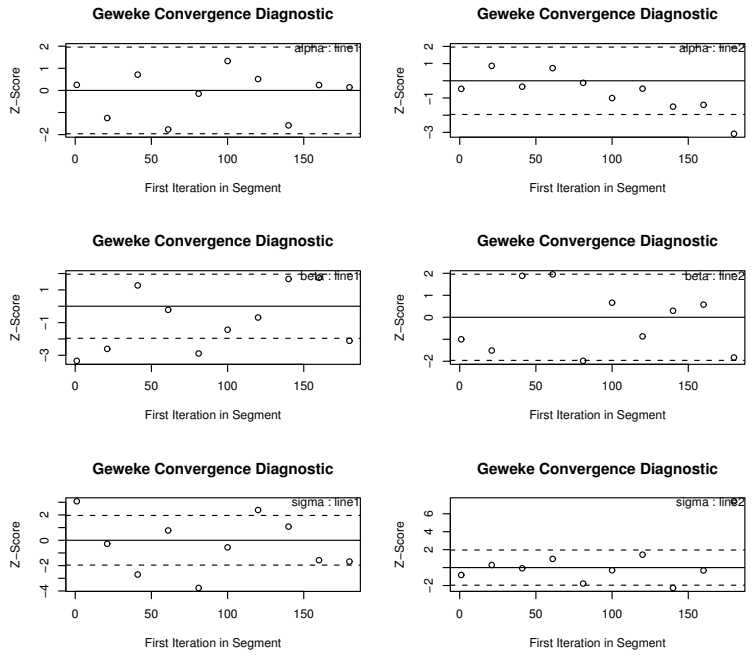
Bayesian Output Analysis








```
6)  Number of Bins:      20
7)  Window Fraction:     0.5

Geweke
------
8)  Alpha Level:         0.05
9)  Number of Bins:      10
10) Window 1 Fraction:   0.1
11) Window 2 Fraction:   0.5

Graphics
--------
12) Device Name:         "X11"
13) Keep Previous Plots: FALSE
14) Plot Layout:         c(3, 2)
15) Separate Plot/Chain: FALSE
16) Separate Plot/Var:   TRUE
```




```
17) Title:                    "Bayesian Output Analysis"
```

The options grouped under the Graphics heading control the general layout used to generate plots. The following gives a brief description of each of these options:

12) The device name used by S-PLUS or R to open graphics windows. Users should not modify this value.
13) If set to "TRUE" all plots generated in BOA will be kept open; otherwise, a value of "FALSE" indicates that only the most recently opened plots be kept open.
14) The number of rows and columns, respectively, of plots to display in one graphics window.
15) If set to "TRUE" only one chain is displayed per plot; otherwise, a value of "FALSE" forces all of the chains to be displayed on the same plot.
16) If set to "TRUE" only one parameter is displayed per plot; otherwise, a value of "FALSE" forces all of the parameters to be displayed on the same plot.
17) The title to be displayed centered at the top of all graphics windows.

Chapter 7

Options Menu

The Options Menu serves as a central location from which the options in Sections 3.1.5, 5.3, and 6.3 can be accessed.

```
GLOBAL OPTIONS MENU
===================
1: Analysis...
2: Data...
3: Plot...
4: All...
5: Return to Main Menu
Selection:
```

Chapter 8

Window Menu

The Window Menu allows the user to switch between and save the active graphics windows.

```
WINDOW 1 MENU
=============
1:Previous
2:Next
3:Save to Postscript File
4:Close
5:Close All
6:Return to Main Menu
Selection:
```

The number of the active graphics window is displayed in the title of his menu. In this example, graphics window 1 is the active window.

8.1 Previous Graphics Window

Make the previous graphics window in the list of open windows the active graphics window.

8.2 Next Graphics Window

Make the next graphics window in the list of open windows the active graphics window.

8.3 Save to Postscript File

Saves the active graphics window to a postscript file. The user is prompted to enter the name of the postscript file in which to save the contents of the graphics window.

```
Enter name of file to which to save the plot [none]
1:
```

Only the name of the file should be given. The file will be automatically saved in the Working Directory (see Section 3.1.5). Microsoft Windows users may save the graphics window in other formats directly from the S-PLUS or R program menus.

8.4 Close Graphics Window

Close the active graphics window.

8.5 Close All Graphics Window

Closes all open graphics windows.

Chapter 9

S-PLUS and R Basics

9.1 Output Display Options

The *options* function in S-PLUS and R can be used to control the format of the outputted text in BOA. This should be done prior to starting BOA. To set the number of significant digits to be displayed, type

```
options(digits = <value>)
```

The number of characters allowed per line can be controlled by entering the command

```
options(width = <value>)
```

9.2 Vectors in S

Several menu selections in BOA prompt the user to input a vector of data. Vectors in S can be supplied in a variety of ways. The simplest way to construct a vector is with the concatenation function *c*:

```
c(<element 1>, <element 2>, ..., <element n>)
```

where the elements may be numerical or logical values or character strings. Another means of constructing vectors is with the *seq* function:

```
seq(<starting value>, <ending value>, length = <number of values>)
```

or

```
seq(<starting value>, <ending value>, by = <step size>)
```

where "length" is number of values in the vector and "by" is the spacing between successive values in the vector. The ":" operator, which is a special case of the *seq* function, can also be used to construct vectors. This operator can be defined as

```
<starting value>:<ending value>
   = seq(<starting value>, <ending value>, by = 1)
```

For more detailed information about these functions, consult the help systems in S-PLUS or R.

9.3 Memory Allocation in R

R and S-PLUS differ in the way they request memory from the operating system. The amount of memory allocated to S-PLUS is free to grow provided the computer's memory resources have not been exhausted. In contrast, a fixed amount of memory is allocated to R when it first starts up. If too little memory is allocated, R may prematurely run out of memory when trying to import large datasets. The amount of allocated memory may be specified by appending the options "–vsize" and "–nsize" to the end of the command used to start the R program. Type "help(Memory)" in R for detailed information on using these options. The following error indicates that the vsize is too small:

```
Error: heap memory (6144 Kb) exhausted [needed 125 Kb more]
       See "help(Memory)" on how to increase the heap size.
```

whereas, the following message is displayed when the nsize is too small:

```
Error: cons memory (250000 cells) exhausted
       See "help(Memory)" on how to increase the number of cons cells.
```

For analyzing extremely large datasets, users may want to experiment with different values for these options until they find a combination that works with their data. On systems with 128Mb of physical memory, a reasonable starting point might be

```
--vsize=64M --nsize=4M
```

Bibliography

[1] Applegate, D., Kannan, R. and Polson, N.G. (1990). Random polynomial time algorithms for sampling from joint distributions. *Technical report no. 500, Carnegie-Mellon University.*

[2] Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7(4)**, 434-455.

[3] Brooks, S.P. and Roberts, G.O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, **8(4)**, 319-335.

[4] Chen, M-H. and Shao, Q-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, **8(1)**, 69-92.

[5] Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883-904.

[6] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-511.

[7] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press.

[8] Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, **31**, 1109-1144.

[9] Jennison, C. (1993). Discussion of "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods." by Smith and Roberts, *Journal of the Royal Statistical Society, Series B*, **55**, 54-56.

[10] Raftery, A. L. and Lewis, S. (1992a). Comment: One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**, 493-497.

[11] Raftery, A. L. and Lewis, S. (1992b). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press.